DOCUMENT RESUME

ED 314 449 TM 014 264

AUTHOR Giroir, Mary M.; Davidson, Betty M.
TITLE Ways of Estimating the Probability That Results Will Replicate.
PUB DATE 9 Nov 89
NOTE 17p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Generalizability Theory; Hypothesis Testing; Probability; Research Design; Sample Size; *Sampling; Scientific Methodology; *Statistical Significance
IDENTIFIERS Bootstrap Hypothesis; *Cross Validation; Jackknifing Technique; *Research Replication; Research Results

ABSTRACT

Replication is important to viable scientific inquiry; results that will not replicate or generalize are of very limited value. Statistical significance enables the researcher to reject or not reject the null hypothesis according to the sample results obtained, but statistical significance does not indicate the probability that results will be replicated. Three techniques for evaluating the sampling specificity of results are described: (1) the jackknife technique of J. W. Tukey (1969); (2) the bootstrap technique of Efron, described by P. Diaconis and E. Bradley (1983); and (3) cross-validation methods described by B. Thompson (1989). A small data set developed by B. Thompson in 1979 is used to demonstrate the cross-validation procedure in detail. These three procedures allow the researcher to examine the replicability and generalizability of results and should be used frequently. Two tables present the study results, and an appendix gives examples of commands for the Statistical Analysis System computer package used for the cross-validation example. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314449

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MARY M. GIROIR
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Ways Of Estimating The Probability That Results Will Replicate

Mary M. Giroir Betty M. Davidson

University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 9, 1989.

TM014264

Abstract

The conclusions in studies are only as valuable as the replicability of study results; results that will not replicate or generalize are of limited value indeed. Statistical significance testing can not be employed to determine if results are replicable or generalizable. The present paper discusses three methods--jackknife, bootstrap, and cross-validation--which can be utilized to determine if results are replicable and generalizable.

The conclusions in studies are only as valuable as the replicability of study results; results that will not replicate or generalize are of limited value indeed (Campo, 1988). As Carver (1987) asserts, "replication is the cornerstone of science" (p. 392). Neale and Liebett (1986, p. 290) explain why replication is important to viable scientific inquiry:

> No one study, however shrewdly designed and carefully executed, can provide convincing support for a causal hypothesis or theoretical statement in the social sciences... . Too many possible (if not plausible) confounds, limitations on generality, and alternative interpretations can be offered for any one observation. Moreover, each of the basic methods of research (experimental, correlational, and case study) and techniques of comparison (within-or between-subjects) has instrinsic limitations. How, then, does social science theory advance through research? The answer is, by collecting a diverse body of evidence about any major theoretical proposition.

Crask and Perreault (1977) assert that failure to determine the validity of sample-based results and the stability of calculated findings may lead to the reporting of inaccurate results based upon sample specific findings.

Unfortunately, far too few researchers attend to replicability issues. In the words of Huck, Cormier and Bounds (1974, p. 369),

> Replication is one of the basic principles of competent research. Yet, if we were to survey the professional journals or textbooks about research in the behavioral sciences, we would find that behavioral scientists have paid little attention to this principle.

One reason why too few researchers evaluate the replicability of their results is the common misperseption, entirely without foundation, that significance test p's inform the researcher regarding the likelihood that results will replicate (Carver, 1978). Statistical significance enables the researcher to reject or not reject the null hypothesis according to the sample results obtained. But statistical significance does not indicate the probability that results will be replicated. Therefore, statistical significance does not reflect the reliability or replicability of results. Even Fisher himself acknowledged this:

> The modern test of significance, before which so many editors of psychological journals are reported to bow down, owes more to R. A. Fisher than to any other man. Yet Sir Ronald's standard of firm knowledge was not one very extremely significant

> result, but rather the ability to repeatedly get results significant at 5%. Repetition is the basis for judging variability and significance and confidence. Repetition of results, each significant, is the basis, according to Fisher, of scientific truth (Tukey, 1969, p. 85).

Relying on statistical significance testing may deceive the researcher in interpretation of the results, as Thompson (1989) illustrates through the use of an example data set. Thompson uses varying sample sizes to demonstrate that the value of the effect sizes remaines unchanged even when the sample size increases. Interpretation of statistical significance, does change, however, as a function of sample size changes. Establishing decisions on statistical significance may influence the researcher to ignore large effect sizes that are not significant while overinterpreting effect sizes that are small but statistically significant. Thus significance testing does not inform the researcher of the likelihood of the replication of results.

Three procedures for evaluating sampling specificity of results have been identified in the literature: a) Tukey's jackknife technique (Crask & Perreault, 1977), b) Efron's bootstrap technique (Diaconis & Bradely, 1983), and c) cross-validation methods (Thompson, 1987, 1989). All three methods generate fake data sets from the original data and

assess the variability over all sets of the fake data to provide empirical estimates of the generalizability of effect sizes (Diaconis & Bradley, 1983). The present paper discusses each procedure and interpretation of results. A small data set develpoed by Thompson (1989) is utilized to demonstrate in detail the cross-validation procedure and how results are interpreted.

Jackknife Procedure

The Jackknife procedure was invented by Maurice Quenoville in 1949 and developed by John Tukey in the 1950's (Diaconis & Bradley, 1983). According to Crask and Perreault (1977), "the essence of the jackknife approach is to partition out the impact of effect of a particular subset of data on an estimate derived from the total sample" (p. 61).

The procedure involves arbitraily deleting one observation at a time from the original sample and then recalculating the effect size for each of the resulting truncated data sets. The researcher is able to determine fluctuations in sampling error which may be contributed to the uniqueness of the single observation deleted or to the combined characteristics of the subset (Campo, 1988; Crask & Perreault, 1977; Diaconis & Bradley, 1983). An example illustrating the application of the jackknife is provided by Crask and Perreault (1977).

Bootstrap Procedure

Efron and his colleagues invented the bootstrap technique in order to estimate the statistical accuracy of a subset from the data in a single sample. The premise, according to Diaconis and Bradley (1983), is to mimic the process of selecting many samples of size N in order to find the probability that the values of the sample statistics will fall within various intervals. The data are copied an enormous number times and a megafile is created. A hugh number of samples of size N are then selected at random and the effect size is calcualted for each sample size to determine how sampling influences effect size. Determiningg replicability of results using the bootstrap technique is especially useful when a large or moderate effect size is obtained, but statistically nonsignificant findings have occurred due to a small sample size (Campo, 1988; Diaconis & Bradley, 1983; Thompson, 1989). Diaconis and Bradley (1983) provide a readable presentation of the bootstrap technique.

Cross-Validation

The cross-validation technique is an elaboration of a simple idea, according to Diaconis and Bradley (1983). The original data are randomly split into subsets. The sample may be split in any porportions, such as thirty percent and seventy percent, or fifty-fifty. The prediction equations

developed for each of the subsets are "crossed" so that each subset uses the other group's prediction equation (Ayabe, 1985; Campo, 1988). The purpose of the cross-validation technique is to evaluate the stability of both effect sizes and the least squares weights (e.g., beta weights) used in all analyses. In order to illustrate this procedure in some depth, an example cross-validation in regression research will be discussed.

Table 1 represents a small data set for a multiple regression problem developed by Thompson (1989). The two predictor variables, "P" and "R", are used to predict the criterion variable, "DV". The first three participants were randomly assigned to the first invariance group ("INV" = "1") and the remainding participants were assigned to the second invariance group ("INV" = "2").

INSERT TABLE 1 ABOUT HERE

Two computer runs must be performed by the researcher conducting an invariance or replicability analysis. The first run establishes the means and standard deviations for the original sample and the two subsets. The subsets may also be referred to as invariance groups. Multiple regression (REG) and the correlation (CORR) procedures are also run for the original data and the invariance groups (Campo, 1988; Thompson, 1989). Appendix A presents the SAS

program implemented to conduct these example analyses for the Thompson (1989) data.

As presented in Table 1, new variables, Z scores (ZP1, ZR1, ZP2 and ZR2) and YHAT predicted scores, are created by the researcher in the second run. A set of YHAT values (YHAT11 and YHAT22) are created by using the invariance group's data and each group's own beta weights. Another set of YHAT values (YHAT12 and YHAT21) is created by using the invariance group's data and the other group's beta weights. By then running the CORR procedure, the invariance results are obtained (Campo, 1988; Thompson, 1989). These results are presented in Table 2.

INSERT TABLE 2 ABOUT HERE

Interpretation of the cross-validation results invokes a comparison of the multiple correlation coefficients, "shrunken" multiple correlation coefficients, and cross-validation or invarianvce correlation coefficients. Krus and Fuller (1982) state that the multiple correlation coefficients are derived from a product-moment correlation between each invariance's predicted scores obtained using its own weights with the criterion scores. According to Campo (1988), the "shrunken" mutliple correlation coefficients are derived using the product-moment correlation between predicted scores of the invariance groups derived using the other group's weights and criterion

scores. The cross-validation coefficient signifies the product-moment correlations between the predicted scores of the invariance groups when each group's own weights are applied as against when the other group's prediction equation is applied (Campo, 1988).

The cross-validation or invariance coefficient represents the correlation between two sets of predicted scores, while the multiple correlation coefficient and the "shrunken R" represent correlations between the invariance group's predicted scores and criterion scores. If the multiple R coefficients are comparable the researcher has evidence for replicability of results (Krus & Fuller, 1982). As Thompson (1989) notes, the researcher "hopes that invariance coefficients will approach one" (p. 4). The invariance coefficients can be directly interpreted aganist the standard of how close to one they are.

For the example analyses reported in Table 2 the multiple correlation coefficients (1.0000 and .8747) and large and comparable. However, in cross-validation both R's "shrink" to negative values (-.2842 and -.5182, respectively). Theoretically, negative values of R cannot occur, so negative values for shrunken R's suggest very poor replicability. Very poor replicability is confirmed by the invariance coefficients (-.2843 and -.5924) reported in Table 2.

Even when favorable replicability results do occur, as as Thompson (1987) notes, "cross-validation results must be interpreted with some caution" (p. 29). Because one data set is split into subsamples, the parameter estimates of the subsamples are interdependent and the results are "liberal" estimates of invariance or replicability (Thompson, 1987).

And when favorable replicability evidence surfaces, the pooled analysis involving all the subjects is still the ultimate basis for deriving the most generalizable interpretations.

As Campo (1988) asserts, "the multiple correlation coefficient for the total group serves as the basis for ultimate interpretation if the results prove to be stable or invariant" (p. 7).

Thompson (1989) stresses the importance of empirically investigating results replicability. Results that look different based on only subjective non-empirical comparisons can, in fact, be quite similar. Statistical significance testing can not be employed to determine if the results will be replicable or generalizable. The three procedures presented--jackknife, bootstrap, and cross-validation--do allow the researcher to examine the replicability and generalizability of results. More researchers should conduct and report these analyses.

References

Ayabe, C. (1985). Multicrossvalidation and the jackknife in the estimation of shrinkage of the multiple coefficient of correlation. Educational and Psychological Measurement, 45, 445-451.

Campo, S. (1988, November). Alternative logics for estimating whether research will generalize. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY.

Carver, P. R. (1978). The case aganist statistical significance testing. Harvard Educational Review, 48, 378-399.

Crask, M. R., & Perreault, W. D. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-67.

Diaconis, P., & Bradley, E. (1983). Computer intensive methods in statistics. Scientific American, 248 (5), 116-130.

Hucks, S. W., Cormier, W. H., & Bounds, W. G., Jr. (1974). Reading statistics and research. New York: Harper and Row.

Krus, D. J., & Fuller, E. A. (1982). Computer assisted multicrossvalidation in regression analysis. Educational and Psychological Measurement, 42, 187-193.

Neale, J. M., & Liebert, R. M. (1986). Science and behvaior:

An introduction to methods of research (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868).

Thompson, B. (1989). Statistical significance, results importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-5.

Tukey, J. W. (1969). Analyzing data: sanctification or detective work? American Psychologist, 24, 83-91.

TABLE 1
Observed and Latent Variables for Thompson's (1989) Data Set

INV	P	R	DV	ZP1	ZR1	YHAT11	YHAT12	ZP2	ZR2	YHAT21	YHAT22
1	1	3	90	-1	-.1323	.5151	-.87339	.	.	.	.
1	2	6	49	0	1.0595	-1.1525	.30350	.	.	.	.
1	3	1	93	1	-.9269	.6370	.57000	.	.	.	.
2	3	8	20	.	.	.	.	-1.162	.669	-.296	-.779
2	5	4	3	.	.	.	.	- .387	-.304	.474	-.411
2	6	0	39	.	.	.	.	.387	-1.276	1.245	-.042
2	7	9	63	.	.	.	.	1.162	.912	-1.423	1.232

TABLE 2
Invariance Statistice

	DV	YHAT11	YHAT12	YHAT21
YHAT11	1.0000[a] (n=3)			
YHAT12	-.2842[b] (n=3)	-.2843[c] (n=3)		
YHAT21	-.5182[b] (n=4)	.	.	
YHAT22	.8747[a] (n=4)	.	.	-.5924[c] (n=4)

[a] The multiple correlation coefficient R for the invariance group.

[b] The "shrunken R" for the invariance group.

[c] The invariance coefficient for the invariance group.

APPENDIX A: Example SAS Commands for Table 1 Data

```
DATA  INVAR;
      INFILE INV
      INPUT INV 1-2 P 4-5 R 7-8 DV 10-11;

if inv=1 then do;
     zp1=(p-2.0)/1.0;
     zr1=(r-3.333)/2.517;
     yhat11=(-.371189*zp1)+(-1.087694*zr1);
     yhat12=(.83549*zp1)+(.286434*zr1);
End;
Else Do;
     zp2=(p-5.5)/1.291;
     zr2=(r-5.25)/4.113;
     yhat21=(-.371189*zp2)+(-1.087694*zr2);
     yhat22=(.83549*zp2)=(.286434*zr2);
End;
PROC PRINT;
PROC MEANS;
     VAR P R DV;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR ALL DATA';
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION USING ALL DATA';


DATA TEMP1;
     SET INVAR;
     IF INVAR=1;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR SUBGROUP ONE';
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION FOR SUBGROUP ONE';


DATA TEMP2;
     SET INVAR;
     IF INVAR=2;
PROC CORR;
     VAR P R DV;
     TITLE1 'DESCRIPTIVE STATISTICS FOR SUBGROUP TWO';
PROC REG;
     MODEL DV=P R/STB;
     TITLE1 'REGRESSION FOR SUBGROUP TWO';
```

```
Data All;
     Set Invar;
Proc Corr;
     Var DV YHAT11 YHAT12 YHAT21 YHAT22;
     TITLE1 'INVARIANCE RESULTS';
```

Note. The analysis requires two runs. The first run includes procedures typed in all capital letters. The second run procedures are typed in bold type. The newly created variables for the second run are typed in lower case.